CSCE 670 - Information Storage and Retrieval

Week 14: Information Retrieval for Science and Scientific Research

Yu Zhang

yuzhang@tamu.edu

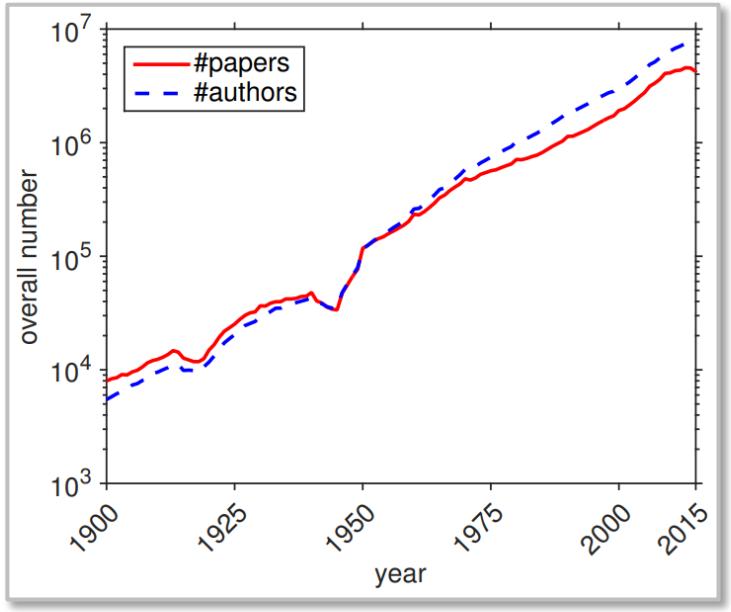
Course Website: https://yuzhang-teaching.github.io/CSCE670-S26.html

Quiz 4!

- All policies are the same as Quiz 1 (number of questions, time limit, grading, etc.)
- Scope:
  - Week 11 (Neural Collaborative Filtering, Sequential Recommendation)
  - Week 12 (LLMs Basics, LLMs with Search Engines)
  - Week 13 (LLMs for Ranking and Recommendation)
  - Homework 3

Explosion of Scientific Text Data

- The volume of scientific publications is growing exponentially.
  - Doubling every 12 years [1]
  - Reaching 240,000,000 in 2019 [2]



- Papers on emerging topics can be released in a torrent.
  - About 4,000 peer-reviewed papers on COVID-19 by the end of April 2020 [3]
  - DeepSeek-R1 was uploaded to arXiv on January 22, 2025 and cited 578 times on April 8, 2025.

Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning [PDF] arx

D Guo, D Yang, H Zhang, J Song, R Zhang... - arXiv preprint arXiv ..., 2025 - arxiv.org

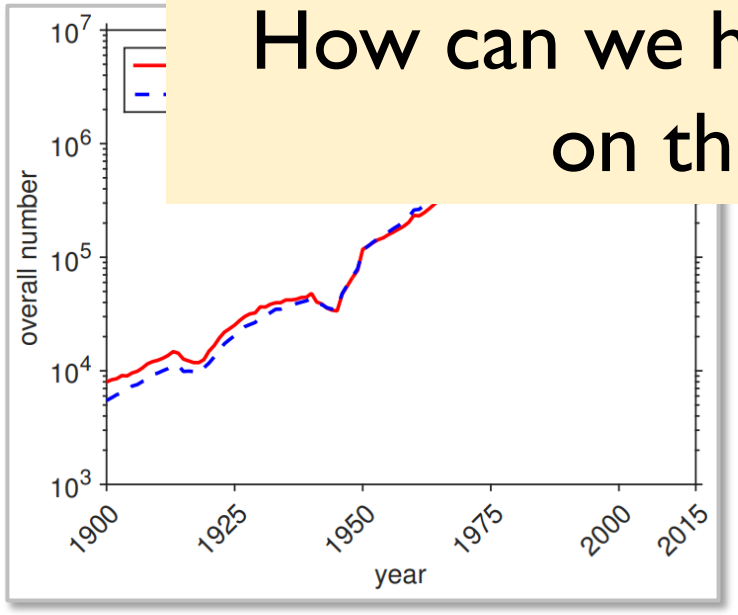
... **DeepSeek-R1**, which incorporates multi-stage training and cold-start data before RL. **DeepSeekR1** ... , we open-source **DeepSeek-R1**-Zero, **DeepSeek-R1**, and six dense models (1.5B, ...

☆ Cited by 578 Related articles »

[1] *A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations*. KDD 2017.
[2] *Microsoft Academic Graph: When Experts are Not Enough*. Quantitative Science Studies 2020.
[3] https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent

Explosion of Scientific Text Data

- The volume of scientific publications is growing exponentially.
  - Doubling every 12 years [1]
  - Reaching 240,000,000 in 2019 [2]



- Papers on emerging topics can be released in a torrent.
  - About 4,000 peer-reviewed papers on COVID-19 by the end of April 2020 [3]
  - [illegible]rXiv on January [illegible]April 8, 2025.

How can we help researchers stay updated on their fields of interest?

[illegible]ability [PDF] arx

[illegible] via reinforcement learning

D Guo, D Yang, H Zhang, J Song, R Zhang... - arXiv preprint arXiv ..., 2025 - arxiv.org

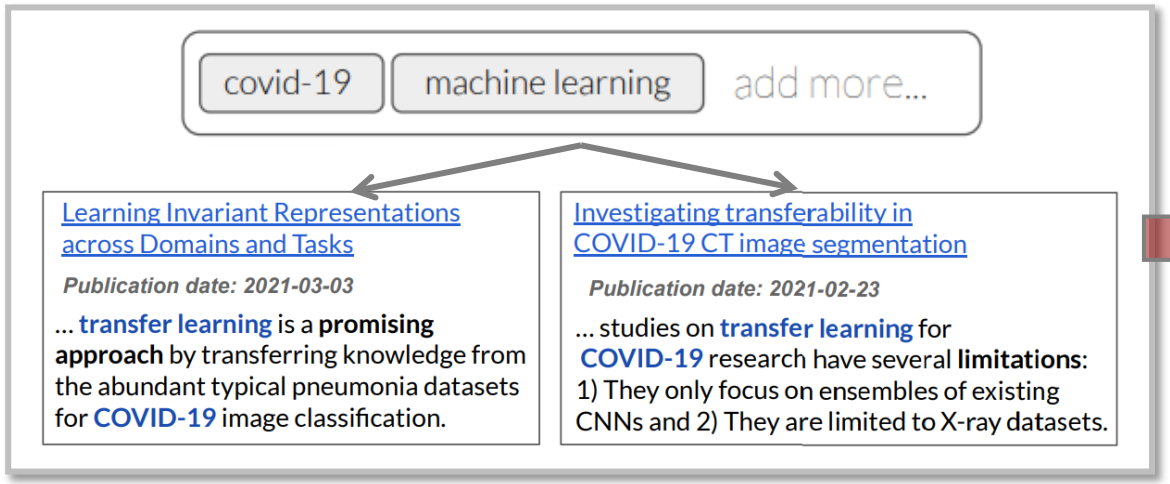
... **DeepSeek-R1**, which incorporates multi-stage training and cold-start data before RL. **DeepSeekR1** ... , we open-source **DeepSeek-R1**-Zero, **DeepSeek-R1**, and six dense models (1.5B, ...

☆ Cited by 578 Related articles »

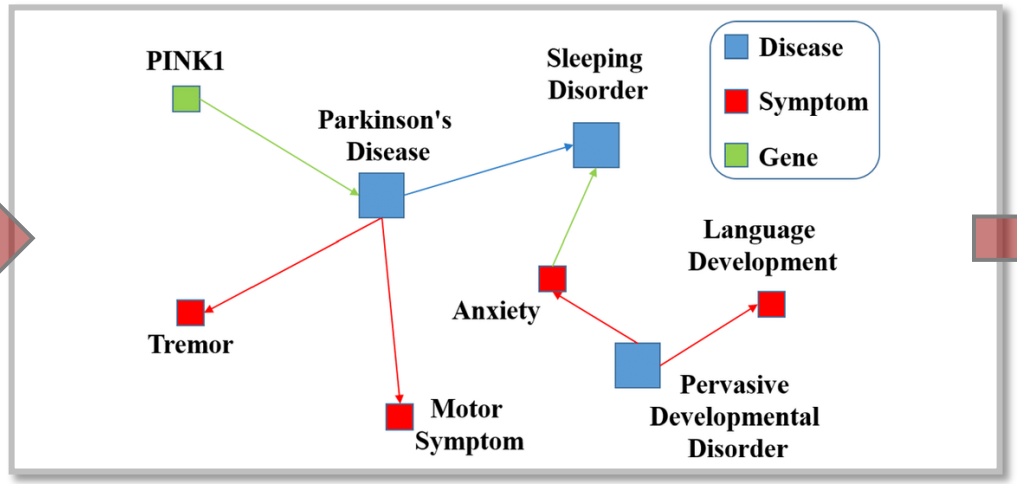
[1] *A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations*. KDD 2017.
[2] *Microsoft Academic Graph:When Experts are Not Enough*. Quantitative Science Studies 2020.
[3] https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent

How can IR assist scientific research?

Finding Relevant Literature



Understanding Literature

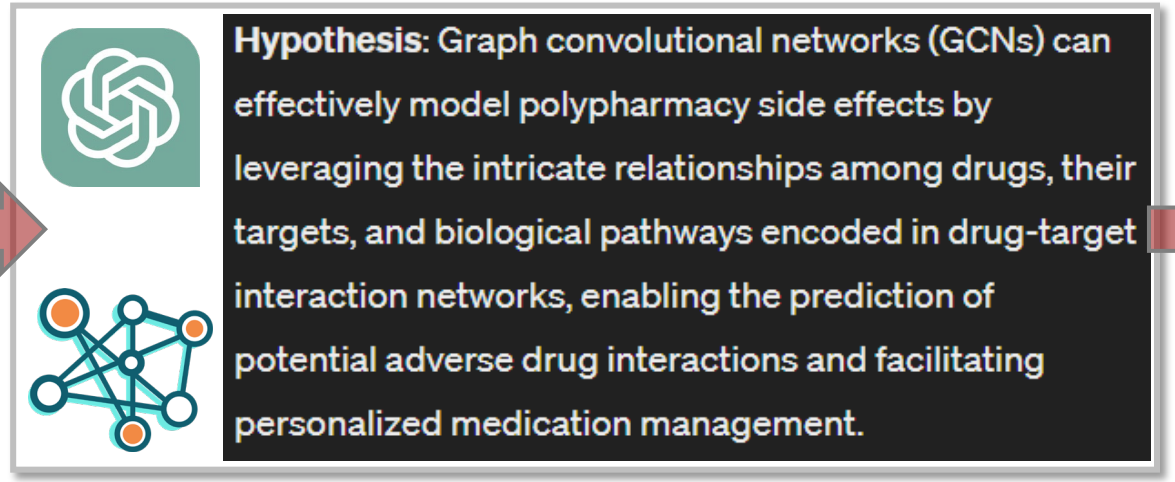


- Example tasks:
  - "*Retrieve top-cited papers relevant to both Betacoronavirus and Paxlovid.*"
  - "*Find papers published in Nature refuting the claim CX3CR1 impairs T cell survival.*"

- Example tasks:
  - "*Which protein is relevant to Parkinson's disease?*"
  - "*What is the relation between Tremor and Sleeping Disorder.*"

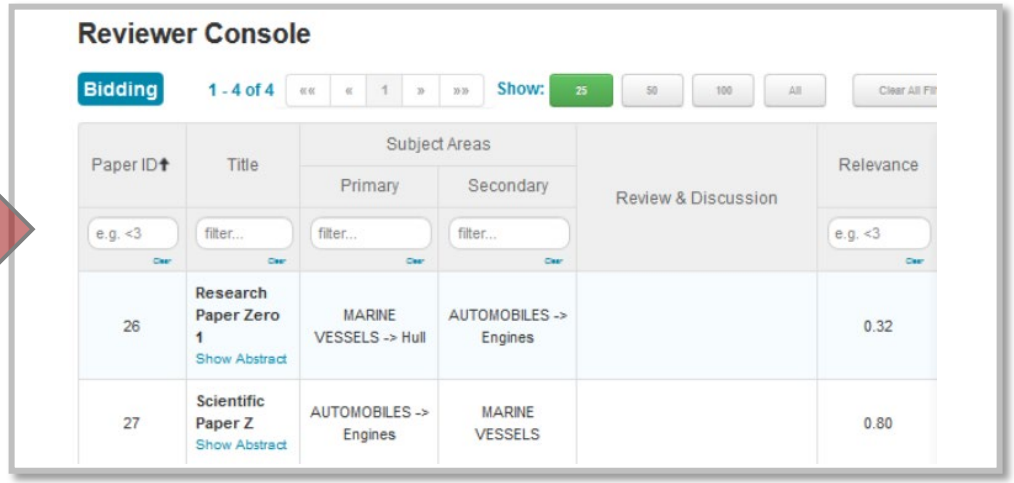
How can IR assist scientific research?

Generating Hypotheses and Suggesting Directions



- Example tasks:
  - "*Generate a new hypothesis on polypharmacy side effects.*"
  - "*Evaluate the novelty of using GNNs for polypharmacy side effect prediction in comparison with previous studies.*"

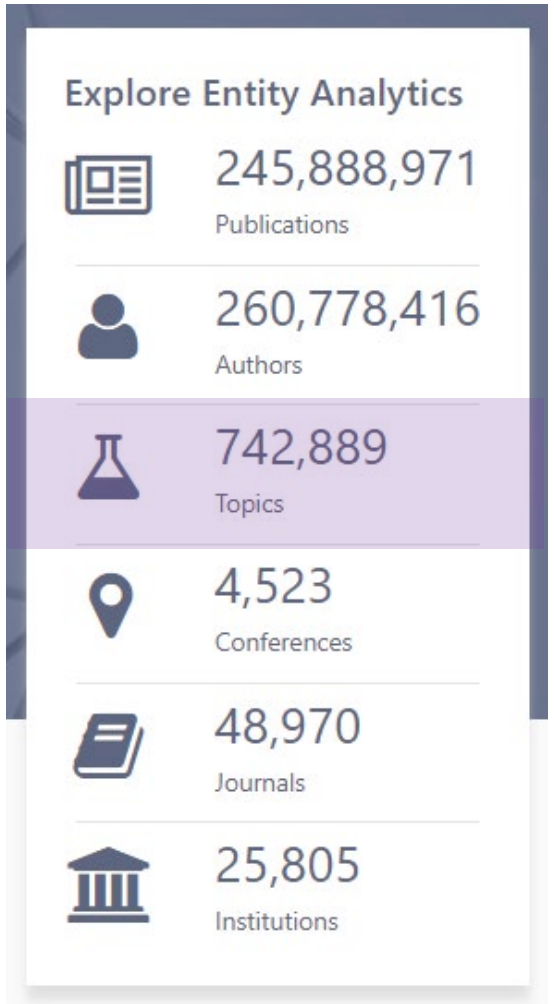
Reviewing Research Outcomes



- Example tasks:
  - "*Provide constructive feedback to the following paper.*"
  - "*Revise the paper according to the following reviews.*"

Literature Retrieval

Fine-Grained Scientific Paper Classification



- The Microsoft Academic Graph has 740K+ categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain 30K+ categories.
- Each paper can be relevant to more than one category (5-15 categories for most papers).

Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.

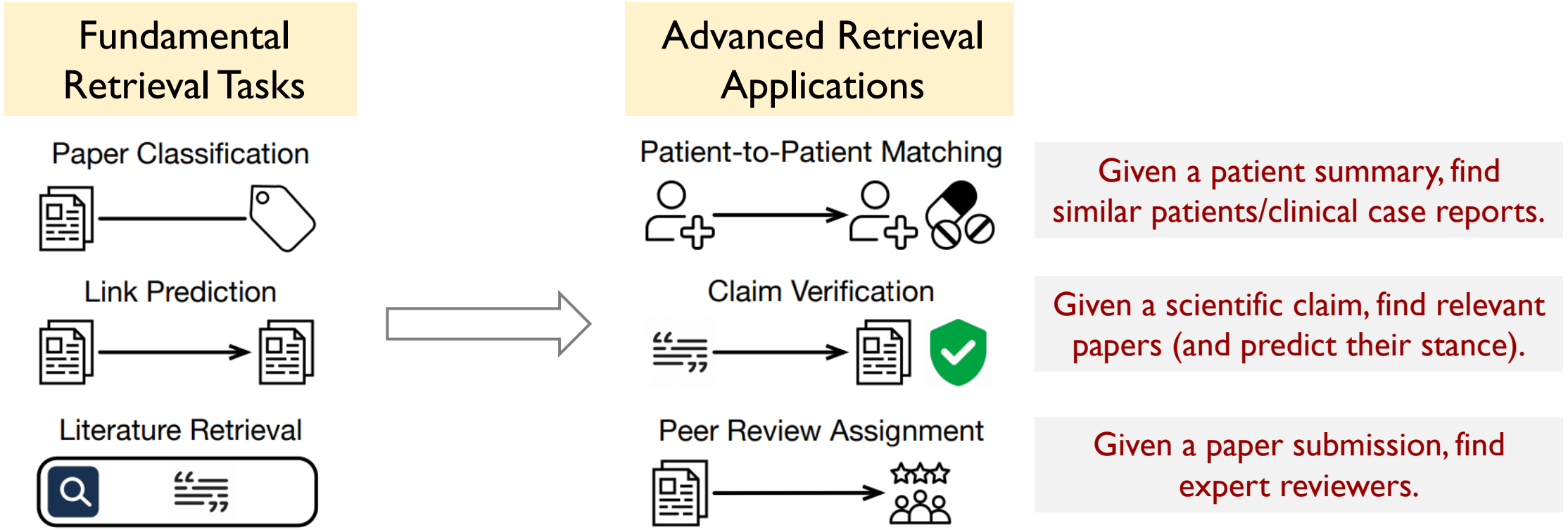
- Relevant categories: Betacoronavirus, Cardiovascular Diseases, Comorbidity, Coronavirus Infections, Fibrin Fibrinogen Degradation Products, Mortality, Pandemics, Patient Isolation, Pneumonia, …

Fine-grained classification can be viewed as a retrieval task.
Query: Paper; Candidates: Category Names

Link Prediction

Link prediction can be viewed as a retrieval task.
Query: Paper; Candidates: Papers

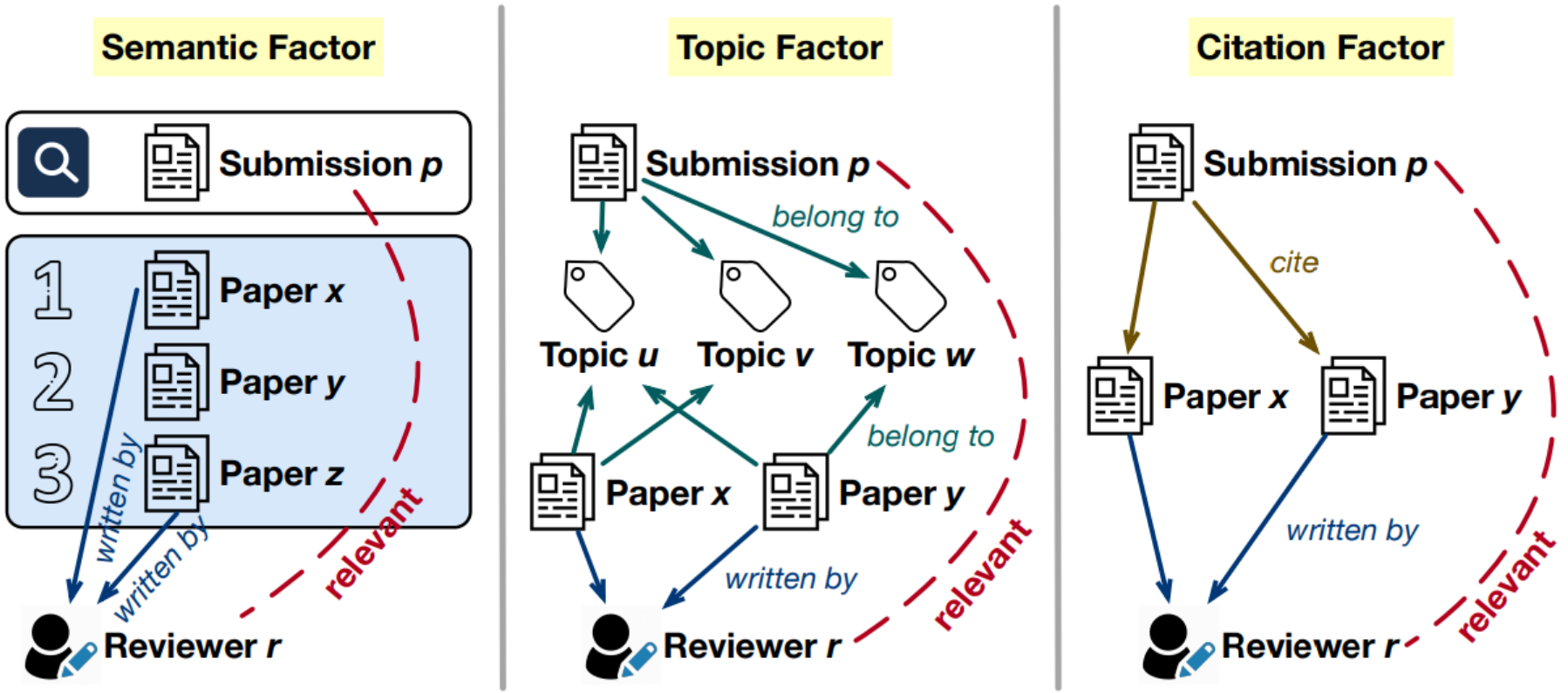
Fundamental Retrieval Tasks vs. Advanced Retrieval Applications



- Why are some tasks more complex?
  - Multiple factors should be considered when judging the relevance.

Multiple Factors for Judging Relevance

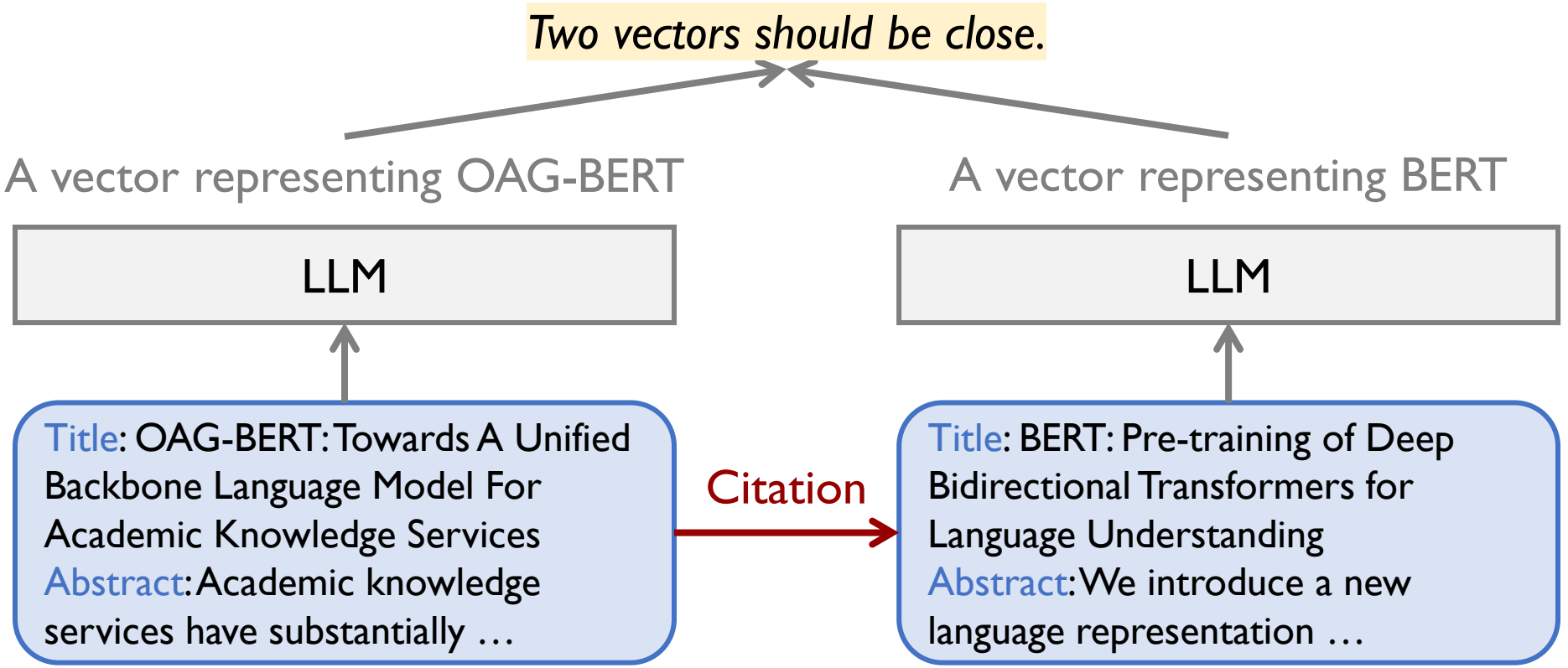
- Example: Paper-Reviewer Matching
  - Why is a pair of (Paper, Reviewer) relevant?



- Multiple factors exist in other tasks (e.g., Patient-to-Article Matching) as well.

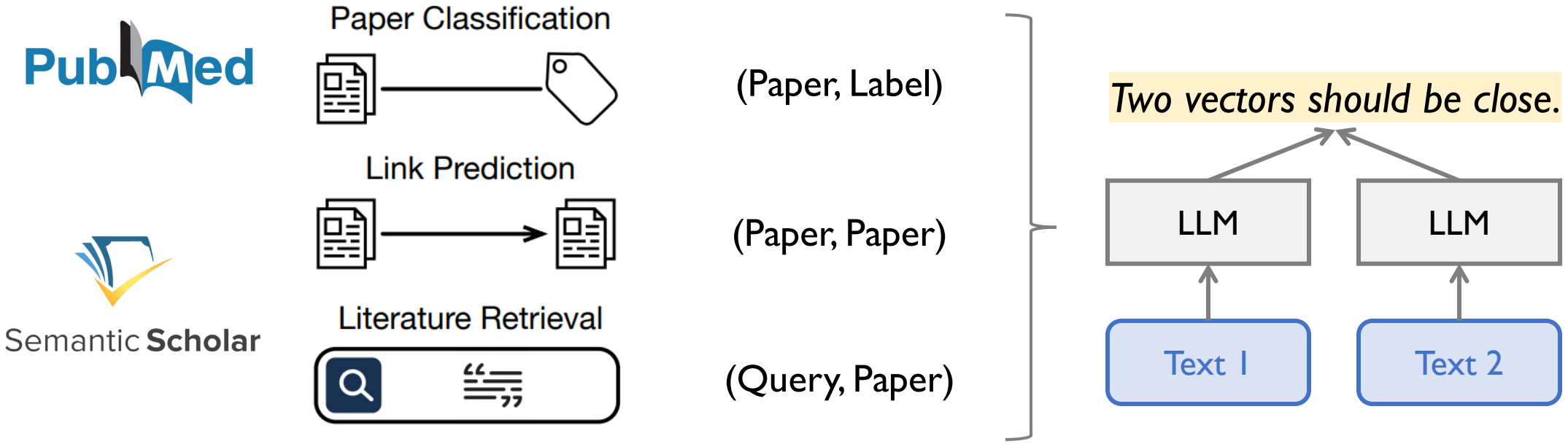
Contrastive Learning for a Fundamental Task

- E.g., Link Prediction
  - Step 1: Collect a large number of papers with citation information.
  - Step 2: Train an LLM with such citation information.



Contrastive Learning for an Advanced Task – A Naïve Way

- Each factor (topic, citation, and semantic) relies on one fundamental retrieval task.
- Directly combining pre-training data from different tasks to train a model?

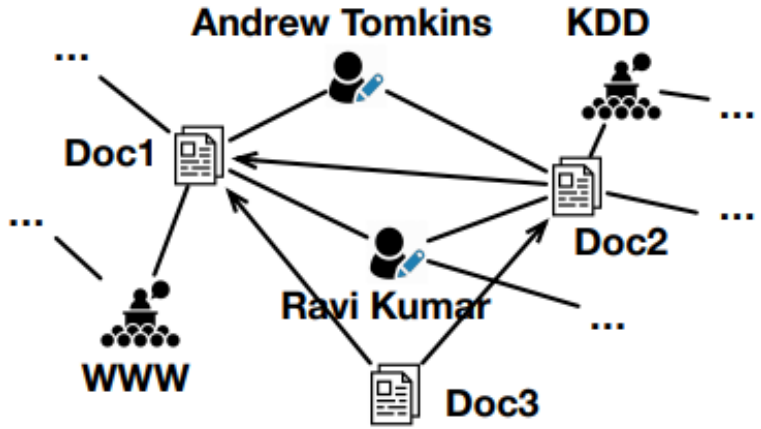


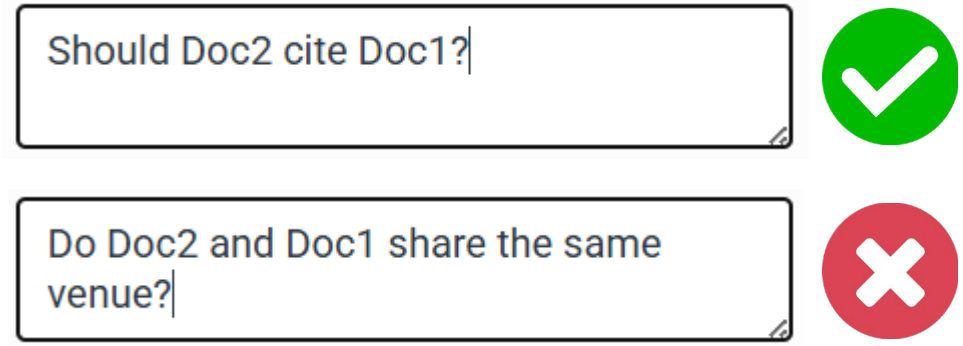
- Task Interference: The model is confused by different types of "relevance".

A Toy Example of Task Interference

- Imagine you have two "tasks".
  - Task 1: Given Paper1 and Paper2, predict if Paper1 should cite Paper2.
  - Task 2: Given Paper1 and Paper2, predict if Paper1 and Paper2 share the same venue.

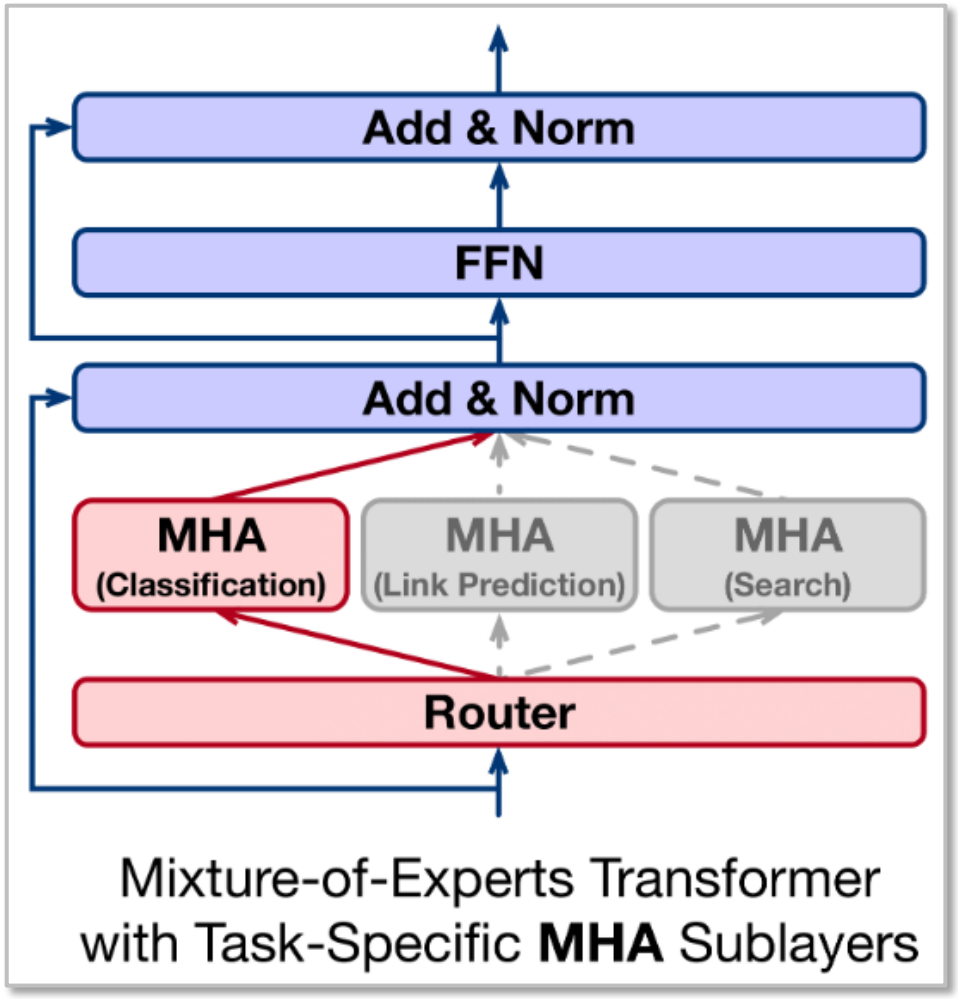
- What if we directly merge the collected relevant (paper, paper) pairs for these two tasks?
  - Is (Doc2, Doc1) relevant?
  - The model does not know which task you are referring to, so it will get confused!





Tackling Task Interference: Mixture-of-Experts Transformer

- A typical Transformer layer
  - 1 Multi-Head Attention (MHA) sublayer
  - 1 Feed Forward Network (FFN) sublayer
- A Mixture-of-Experts (MoE) Transformer layer
  - Multiple MHA sublayers
  - 1 FFN sublayer
  - (Or 1 MHA & Multiple FFN)
- Specializing some parts of the architecture to be an "expert" of one task
- The model can learn both commonalities and characteristics of different tasks.



Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding. EMNLP 2023. (Our work)

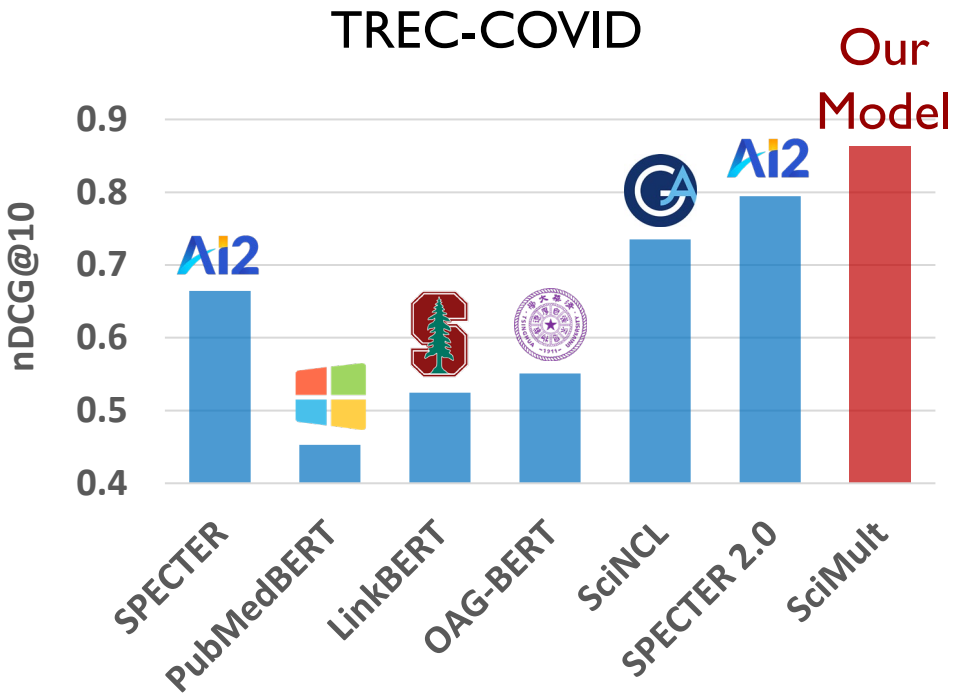
Comparison with Previous Approaches

- New SOTA on the PMC-Patients benchmark (patient-to-article retrieval)
- Outperforming previous scientific pre-trained language models in classification, link prediction, literature retrieval (TREC-COVID), paper recommendation, and claim verification (SciFact)

Patient-to-Article Retrieval (PAR) Leaderboard

	Model	MRR (%)	P@10 (%)	nDCG@10 (%)	R@1k (%)
Our Model 1 June 25, 2023	DPR (SciMult-MHAExpert) (Zhang et al. 2023)	29.89	9.35	13.79	53.71
2 Apr 5, 2023	RRF (Zhao et al. 2023)	29.86	8.86	13.36	49.45

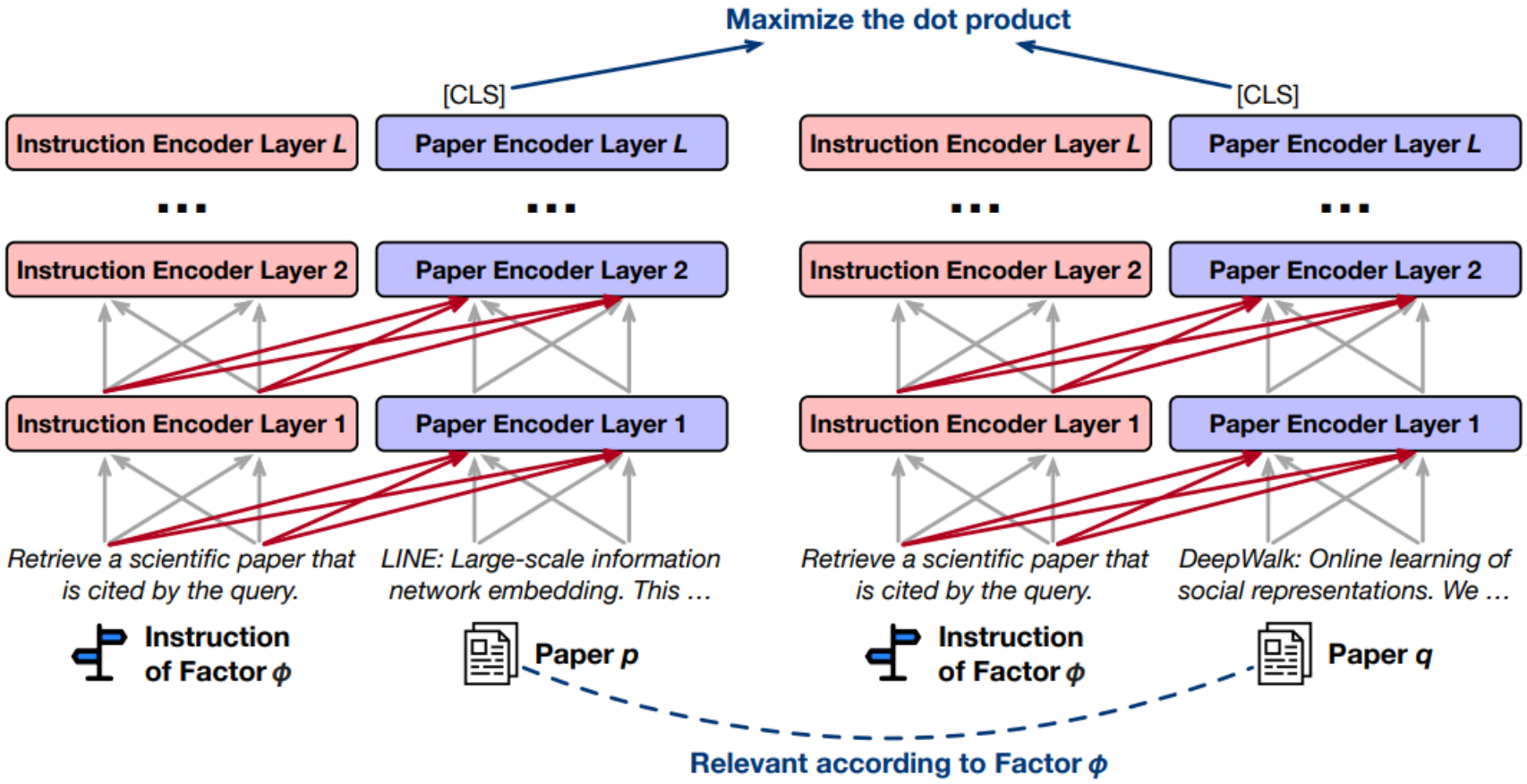
https://pmc-patients.github.io/



Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding. EMNLP 2023. (Our work)

Tackling Task Interference: Instruction Tuning

- Using a factor-specific instruction to guide the paper encoding process
- The instruction serves as the context of the paper.
- The paper does NOT serve as the context of the instruction.



Chain-of-Factors Paper-Reviewer Matching. WWW 2025. (Our work)

Comparison with Previous Approaches

- Public benchmark datasets
  - Expert C judges whether Reviewer A is qualified to review Paper B.
- Outperforming the Toronto Paper Matching System (TPMS, used by Microsoft CMT)

	SciRepEval [44]					SIGIR [19]					KDD				
	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average
TPMS [7]	62.06**	53.74**	31.40**	24.86**	43.02**	39.73**	38.36**	17.81**	17.12**	28.26**	17.01**	16.78**	6.78**	7.24**	11.95**
SciBERT [6]	59.63**	54.39**	28.04**	24.49**	41.64**	34.79**	34.79**	14.79**	15.34**	24.93**	28.51**	27.36**	12.64**	12.70**	20.30**
SPECTER [9]	65.23**	**56.07**	32.34**	25.42	44.77**	39.73**	40.00**	16.44**	16.71**	28.22**	34.94**	30.52**	15.17**	**13.28**	23.48**
SciNCL [34]	66.92**	55.42**	34.02*	25.33	45.42**	40.55**	39.45**	17.81**	17.40*	28.80**	36.21**	30.86**	15.06**	12.70**	23.71**
COCO-DR [56]	65.05**	55.14**	31.78**	24.67**	44.16**	40.00**	40.55*	16.71**	17.53	28.70**	35.06**	29.89**	13.68**	12.13**	22.69**
SPECTER 2.0 CLF [44]	64.49**	55.23**	31.59**	24.49**	43.95**	39.45**	38.63**	16.16**	16.30**	27.64**	34.37**	30.63**	14.48**	12.64**	23.03**
SPECTER 2.0 PRX [44]	66.36**	55.61**	34.21	**25.61**	45.45**	40.00**	38.90**	19.18**	16.85**	28.73**	37.13	31.03	15.86**	13.05*	24.27*
Our Model CoF	**68.47**	55.89	**34.52**	25.33	**46.05**	**45.57**	**41.69**	**22.47**	**17.76**	**31.87**	**37.63**	**31.09**	**16.13**	13.08	**24.48**

: semantic-based method : topic-based method : citation-based method

Chain-of-Factors Paper-Reviewer Matching. WWW 2025. (Our work)

Brainstorming and Long-Horizon Reasoning

OpenResearcher

- A deep research agent capable of 100-step information seeking and reasoning
- https://huggingface.co/spaces/OpenResearcher/OpenResearcher

How can LLMs learn to perform long-horizon research?

- We train LLMs using long-horizon information seeking and reasoning trajectories.
- How to synthesize such trajectories?
- Challenge 1: Number of steps
  - Vanilla RAG: 1 step
  - Graph-CoT / GraphDancer: often 2-5 steps
  - Scientific research: often 20-100 steps
- Challenge 2: No offline environment
  - Most approaches rely on live search APIs
  - Expensive ($1-5 per 1,000 trajectories)
  - Non-deterministic (cannot be reproduced 1 month later)

Method	Price/K	Total Cost
Serper API (Serper.dev, 2026)	$1	$5,760
SerpAPI (Serpapi.com, 2026)	$5	$28,800
Offline retriever (ours)	**$0**	**$0**

OpenResearcher: A Fully Open Pipeline for Long-Horizon Deep Research Trajectory Synthesis. arXiv 2026. (Our work)

Solution: An Offline Corpus + A Retriever + GPT-OSS-120B

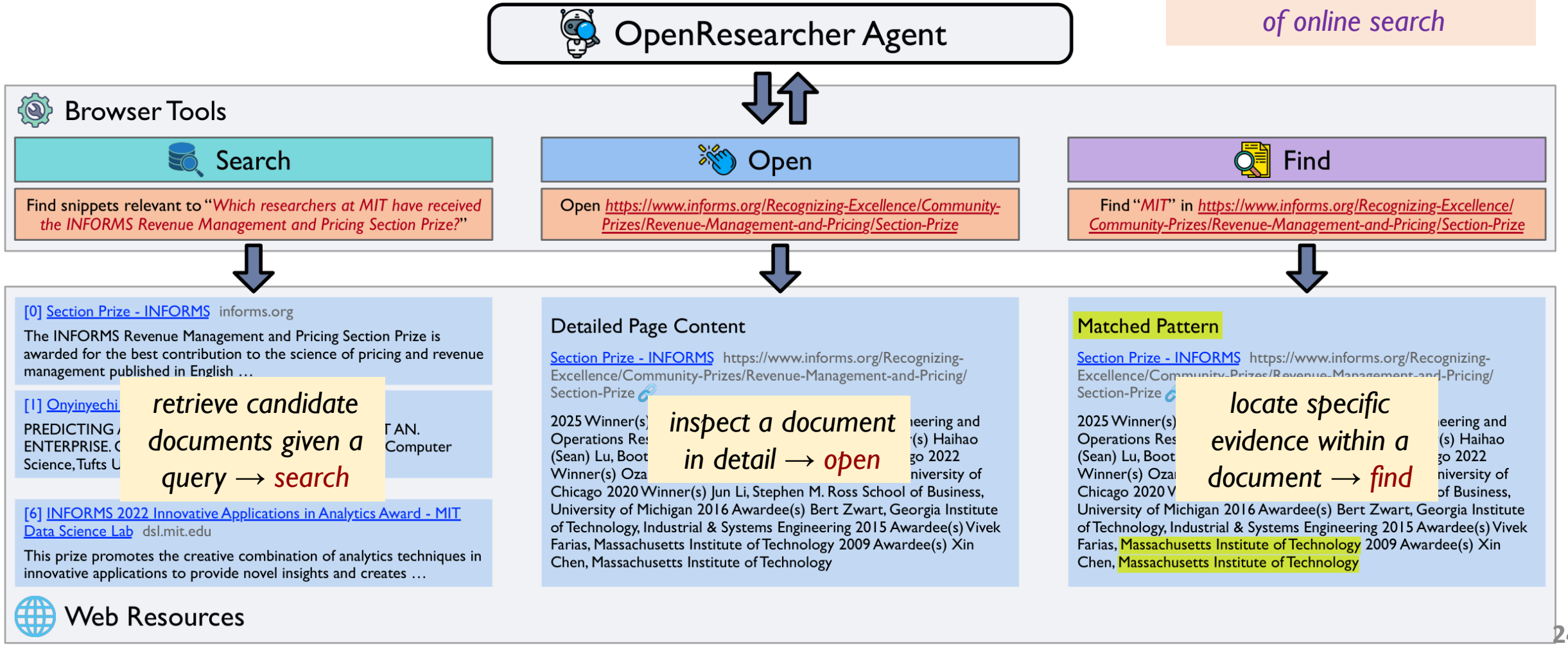
- How to construct the offline corpus?
  - Step 1: Select a set of questions (along with their answers) from a relatively challenging web QA dataset
  - Step 2: Use these questions/answers as queries to retrieve relevant documents from the web
  - Step 3: Then mix in a large number of irrelevant documents to simulate the noise and difficulty of a live web setting
  - Practical Observation: It is crucial that the corpus contains the relevant documents that actually support the answer (Step 2). Otherwise, we cannot tell whether the agent fails because it cannot find them or because such information does not exist.

OpenResearcher: A Fully Open Pipeline for Long-Horizon Deep Research Trajectory Synthesis. arXiv 2026. (Our work)

Solution: An Offline Corpus + A Retriever + GPT-OSS-120B

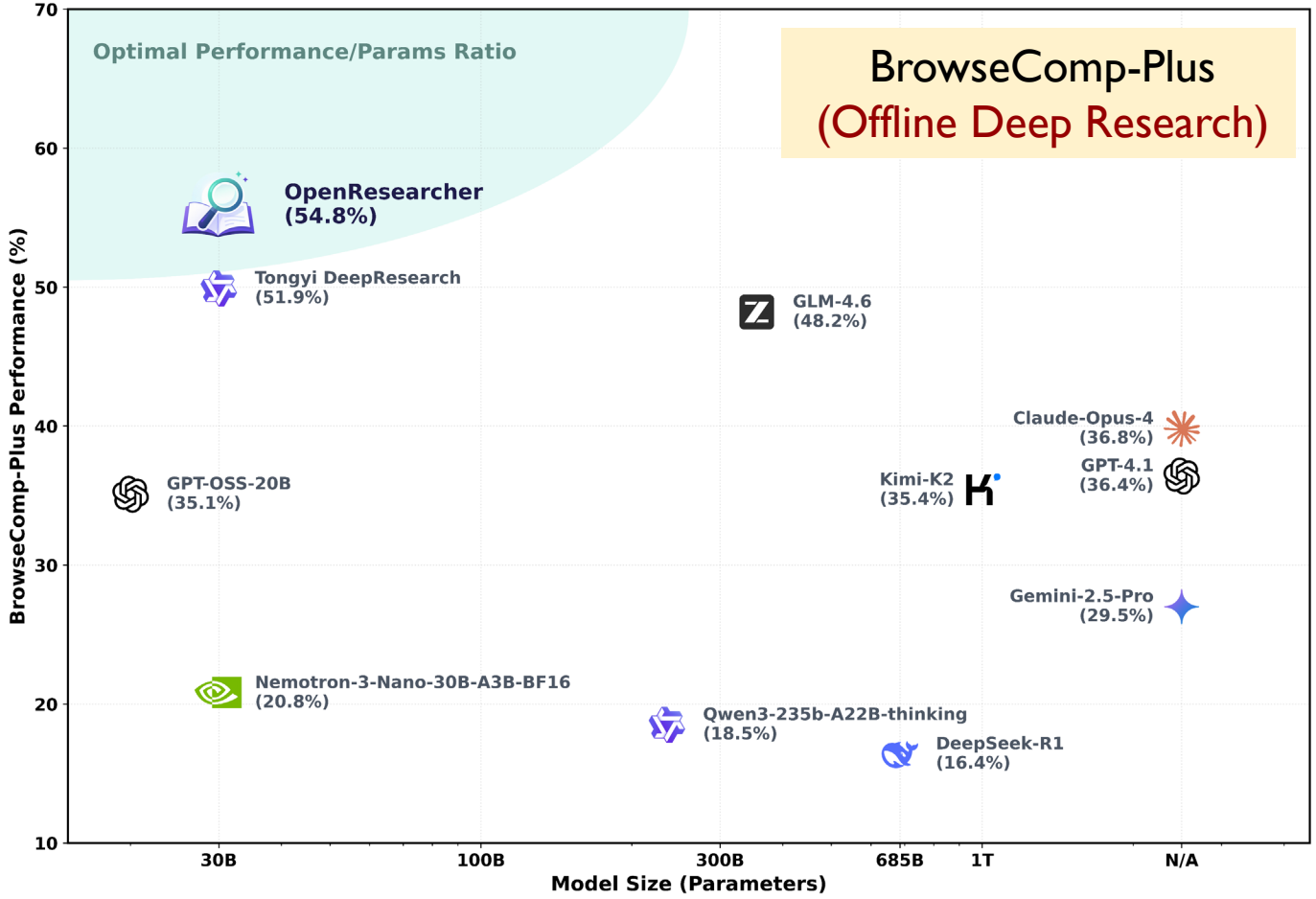
- How to build the retriever?

overall objective: use three tools to simulate the scenario of online search



Performance Comparison

- An Offline Corpus + A Retriever + GPT-OSS-120B → long-horizon trajectories
- Use these trajectories to fine-tune a smaller model: Nemotron-3-Nano-30B-A3B-Base



Online Deep Research

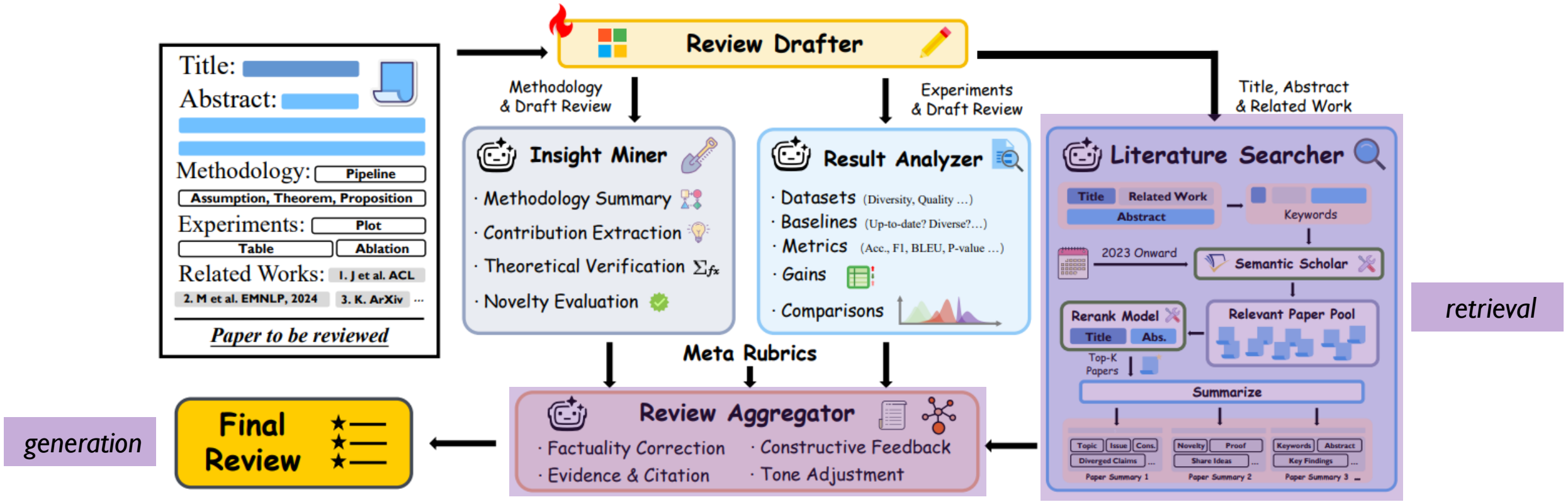
	BrowseComp	GAIA	xbench
Foundation Models with Tools			
OpenAI o4-mini	28.3	55.8	67.0
Claude-4-Sonnet	12.2	57.6	64.0
Kimi-K2	14.1	57.7	50.0
DeepSeek-R1	8.9	30.3	55.0
Nemotron-3-Nano	10.6	50.5	55.0
Deep Research Agents			
ASearcher-QwQ-32B	5.2	52.8	42.0
WebDancer-QwQ-32B	3.8	51.5	39.0
WebSailor-72B	12.0	55.4	55.0
DeepMiner-32B	21.2	54.4	53.0
Ours			
OPENRESEARCHER	**26.3**	**64.1**	**65.0**

Paper Reviewing

How should we evaluate the novelty of a submission?

- The authors claim that their method is novel. ✗
- By searching for and comparing with prior related work, identify the key distinctions of this submission. ✓

ground the review in prior related work!

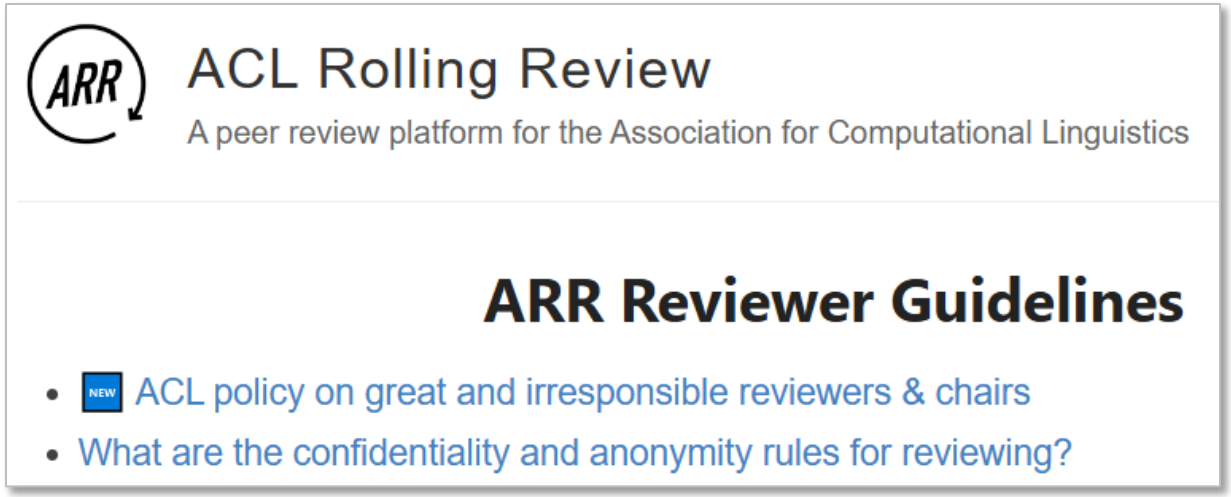


ReviewGrounder: Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents. ACL 2026. (Our work)

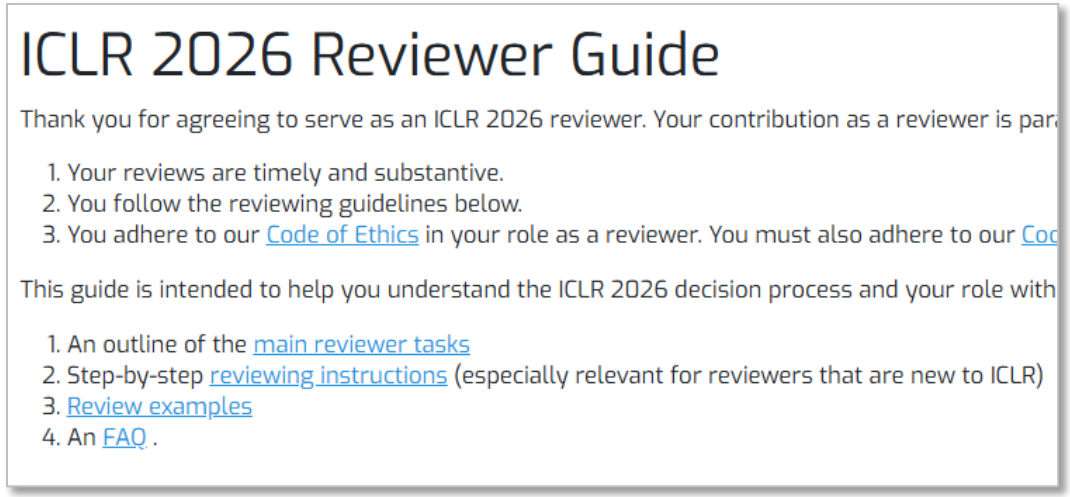
How can we generate substantive critiques and suggestions?

- Superficial and formulaic comments (e.g., "*add more datasets/tasks*") ✗
- Follow official reviewer guidelines that specify what to attend to in different review sections and which criteria to consider across evaluation dimensions ✓

ground the review in official rubrics!



https://aclrollingreview.org/reviewerguidelines



https://iclr.cc/Conferences/2026/ReviewerGuide

ReviewGrounder: Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents. ACL 2026. (Our work)

Performance Comparison

- ReviewGrounder, using a Phi-4-14B drafter and a GPT-OSS-120B grounding stage, consistently outperforms baselines with substantially stronger/larger backbones (e.g., GPT-4.1 and DeepSeek-R1-670B) in both alignment with human judgments and rubric-based review quality.

Method	Model	Core	Res.	Comp.	EBC	Clr.	Cov.	Tone	Contradict.	Overall	Δ
Foundation Model	Qwen3-32B	1.6971	0.7642	0.5800	0.1437	1.6128	1.1537	1.9992	-0.1460	7.8047	↑ 38%
	QWQ-32B	1.6901	0.6531	0.3513	0.1186	1.6792	0.9461	1.9969	-0.0836	7.3517	↑ 46%
	GPT-4o	1.1969	0.1037	0.0302	0.0024	1.0499	0.3318	1.9840	-0.1233	4.5756	↑ 135%
	GPT-4.1	1.7573	0.6966	0.3406	0.1074	1.6327	1.1675	1.9992	-0.0397	7.6616	↑ 41%
AgentReview	GPT-4o	1.1300	0.1600	0.1100	0.1250	1.3400	0.5900	2.0000	-0.1600	4.8675	↑ 121%
	GPT-4.1	1.0300	0.1300	0.1200	0.0000	1.4100	0.6300	1.9800	-0.1600	4.9620	↑ 117%
AI Scientist	GPT-4o	0.8500	0.0000	0.0200	0.0000	0.6700	0.1800	1.7600	-0.1900	3.6800	↑ 193%
	GPT-4.1	1.6700	0.4800	0.3600	0.0830	1.5600	1.1300	1.9400	-0.0900	7.0893	↑ 52%
CycleReviewer	Llama-3.1-8B	0.9852	0.1011	0.0645	0.0089	0.5832	0.1493	1.6571	-0.4504	3.0989	↑ 248%
	Llama-3.1-70B	1.0187	0.1633	0.0980	0.0109	0.7698	0.2551	1.8476	-0.6412	3.5220	↑ 206%
DeepReviewer	Phi-4-7B	1.4207	0.4545	0.3299	0.1311	1.3743	1.0599	1.9432	-0.3953	6.3183	↑ 70%
	Phi-4-14B	1.6306	0.6532	0.4977	0.3532	1.6772	1.2877	1.9930	-0.1922	7.9004	↑ 36%
REVIEWGROUNDER	**Phi-4-14B**	**1.8507**	**1.4075**	**0.9059**	**1.4831**	**1.9191**	**1.3289**	**1.9992**	**-0.1245**	**10.7699**	-

Table 1: **Performance comparison of reviewer models on REVIEWBENCH under rubric-based evaluation.** We visualize gains of REVIEWGROUNDER to each baseline in the Δ columns. Notes: Higher scores indicate better performance; Contradict. is a pitfall dimension scored in $-2, -1, 0$, while others are scored in $0, 1, 2$. *Abbreviations:* Core=CORE CONTRIBUTION ACCURACY, Res.=RESULTS INTERPRETATION, Comp.=COMPARATIVE ANALYSIS, EBC=EVIDENCE-BASED CRITIQUE, Clr.=CRITIQUE CLARITY, Cov.=COMPLETENESS COVERAGE, Tone=CONSTRUCTIVE TONE, Contradict.=FALSE OR CONTRADICTORY CLAIMS.

Method	Model	Decision		Rating	
		ACC↑	F1↑	MSE↓	MAE↓
AgentReview	Claude-3-5-sonnet	0.2826	0.2541	2.8406	1.2989
	Gemini-2.0-Flash-Thinking	0.4242	0.4242	2.6186	1.2170
	DeepSeek-V3	0.3140	0.2506	1.9951	1.1017
AI Scientist	GPT-o1	0.4167	0.4157	4.3072	1.7917
	Claude-3-5-sonnet	0.5579	0.4440	3.0992	1.3500
	Gemini-2.0-Flash-Thinking	0.6139	0.4808	3.9232	1.6470
	DeepSeek-V3	0.4059	0.3988	4.8006	1.8403
	DeepSeek-R1	0.4259	0.4161	4.7719	1.8099
CycleReviewer	Llama-3.1-8B	0.2354	0.3988	3.1324	1.3663
	Llama-3.1-70B	0.1545	0.4156	1.8440	1.0643
DeepReviewer	Phi-4-7B	0.6381	0.6068	1.4442	0.9416
	Phi-4-14B	0.6667	0.5204	1.3527	0.9041
REVIEWGROUNDER	**Phi-4-14B**	**0.6939**	**0.6699**	**1.1607**	**0.8597**

Table 2: **Performance comparison of reviewer models on REVIEWBENCH under numeric-field evaluation.**

ReviewGrounder: Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents. ACL 2026. (Our work)

Logistics for Final Exam

Final Exam (Overall Information)

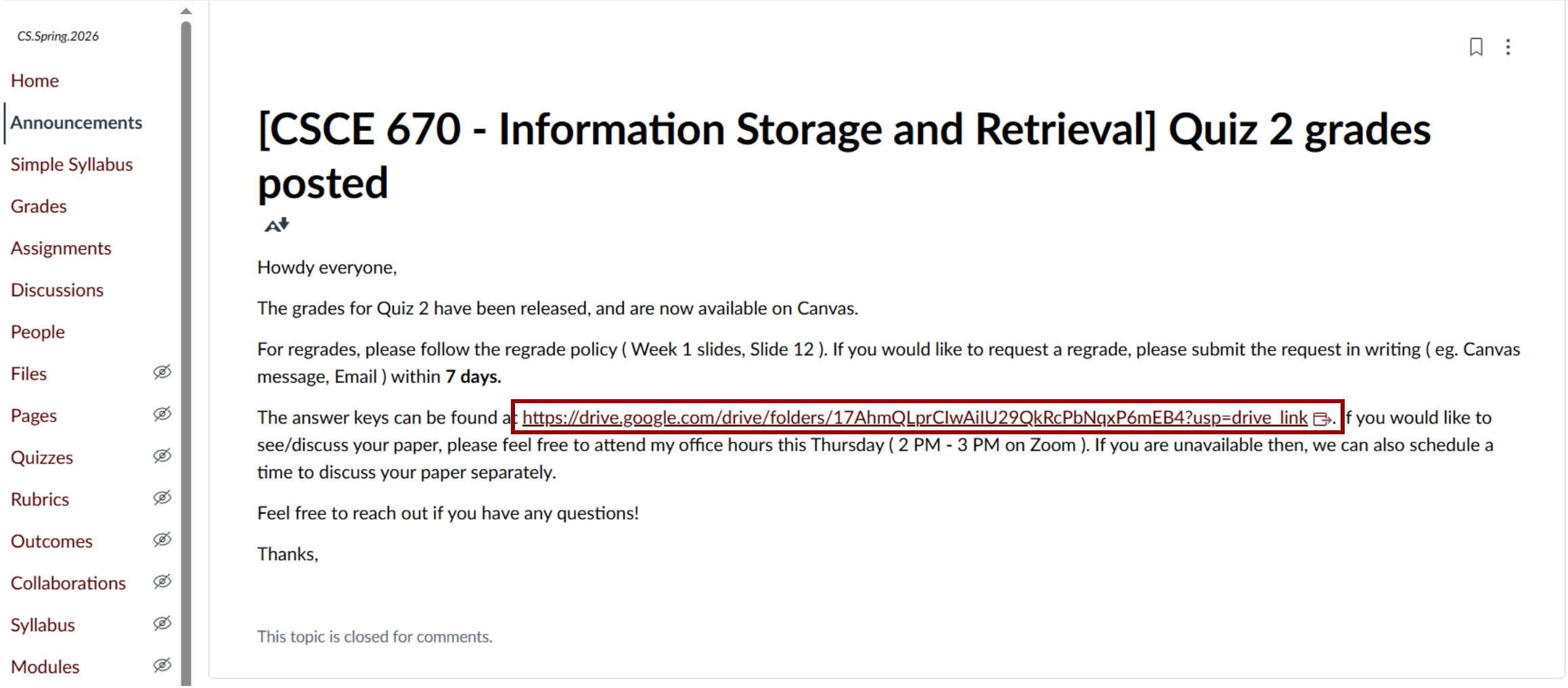
- Time: 3:30pm – 5:30pm on May 4, 2026 (Monday)
- Location: In our regular classroom
- Scope: Comprehensive! Covering all materials from Week 1 to Week 13, EXCEPT FOR
  - content marked as "*will not appear in quizzes or the exam*" (i.e., slides with a gray title)
  - homework assignments

Final Exam (Format)

- 30 multiple-choice questions, 1% each
  - In total, the final exam accounts for 30% of your overall grade.
- 15 of the questions will be VERY SIMILAR to those that appeared in Quiz 1-Quiz 4
  - Test your "few-shot" problem-solving skills
  - Carefully review all quizzes (both versions), but you can skip the last question, as the final exam does NOT cover the homework assignments
- The other 15 questions will not have appeared on any quizzes, but their difficulty will be comparable to the quiz questions
  - Test your "zero-shot" problem-solving skills

Where can you find the quiz questions and answers?

- Canvas → "Announcements" tab → Quiz XXX grades posted → Google Drive link



Reminder: Time Allocation

- Quizzes: 40 minutes, 7 questions
- Final Exam: 120 minutes, 30 questions!

- If you found the quiz time tight, you will need to solve problems faster on the final than you did on the quizzes!
  - Or my recommended strategy: your speed on seen questions should be twice as fast as on unseen questions (i.e., 40 minutes for 15 seen questions; 80 minutes for 15 unseen questions)

Final Exam (Other Details)

- Cheat Sheet: You can bring one cheat sheet
  - 8.5" x 11" standard sheet of paper with anything on it
  - Front and back
  - Either printed or handwritten is okay

- Laptops, books, and notes other than the cheat sheet are NOT allowed
- Calculators are NOT required, and the questions will NOT involve calculations that cannot be done easily by hand.

Final Exam (Conflict)

- Please refer to Student Rule 7 (https://student-rules.tamu.edu/rule07/) and Rule 8 (https://student-rules.tamu.edu/rule08/)

- You need to submit a written request and supporting documentation in advance, unless an unforeseen circumstance occurs that prevents prior notice (e.g., sudden illness on the day of the exam)

Logistics for Project Presentations

Project Presentations (Overall Information)

- Time: Apr 20, Apr 22, Apr 24, and Apr 27 in class
  - For the detailed schedule, please refer to the shared Google Sheet
- Location: Zoom (https://tamu.zoom.us/j/6411788612; you can find it on the course webpage)
- Presenters: Each group may have one presenter, multiple presenters, or the entire group
  - Regardless of how many people present, each group has a total of 9 minutes. (For prototype projects, please plan for 5-6 minutes of presentation and 3-4 minutes of live demo)
  - At 9 minutes, I will unmute myself and notify you, and you will have 1 minute to wrap up everything. If you are not finished after that extra minute, a 1% overtime penalty will be applied.

Project Presentations (Rubrics for a Prototype Project)

- Function of your system and motivation (1%)
- Existing similar systems and their limitations (1%)
- Data (1%)
  - You may include: dataset statistics, collection/annotation process, …
- Implementation (2%)
  - You may include: core algorithm, existing resources used, …
  - If you have unfinished parts, explain how to finish them before the project report due (May 2)
- Live demo (3%)
  - Minimum requirement: Given a sample input (e.g., query or item), show your system's output (relevant documents or recommended items)
- Finish the presentation within the allotted time (1%)

Project Presentations (Rubrics for a Research Project)

- Task background and definition (1%)
- Related work and their limitations (1%)
- Proposed solution (2%)
  - You may include: model architecture, objective function, …
- Data (1%)
  - You may include: dataset statistics, collection/annotation process, …
- Experimental results (2%)
  - You may include: metric, comparison with baseline(s), ablation study, case study, …
- Take-away messages (1%)
  - If you have unfinished parts, explain how to finish them before the project report due (May 2)
- Finish the presentation within the allotted time (1%)

Project Presentations (Rubrics for a Survey Project)

- Survey scope and motivation (1%)
- Existing related surveys and their limitations (1%)
- Survey structure and approach (2%)
  - You may include: organizing principle (by tasks, methods, applications, …)
- Representative papers (2%)
- Insights and take-away messages (2%)
  - You may include: commonalities and characteristics of different works, lessons for readers, …
  - If you have unfinished parts, explain how to finish them before the project report due (May 2)
- Finish the presentation within the allotted time (1%)

Office Hours

- The remaining office hours are scheduled as follows:
  - Additional office hour on May 3; https://tamu.zoom.us/j/6411788612

Sun	Mon	Tue	Wed	Thu	Fri	Sat
	Apr 20 Yu (3:30-4:30 pm, PETR 222)			Apr 23 TA (2-3 pm, EABC)		
	Apr 27 Yu (3:30-4:30 pm, PETR 222)			Apr 30 TA (2-3 pm, Zoom)		
May 3 Yu (3-5 pm, Zoom)	May 4 Final (3:30-5:30 pm, HRBB 113)					

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE670-S26.html